Compliance Data Warehouse (CDW) – Privacy Impact Assessment

PIA Approval Date – Mar. 24, 2008

System Overview:
The Compliance Data Warehouse (CDW) is an official system as identified in the IRS As-Built Architecture, and contains sensitive but unclassified (SBU) taxpayer data from various internal data sources. The purpose of CDW is to provide high quality data and information services, primarily to the Research community, in support of projects, analyses, and studies related to tax administration, enforcement, and customer service. CDW is also accessible to the Department of Treasury, Office of Tax Policy, Office of Tax Analysis – via an approved Interconnection Security Agreement and Memorandum of Understanding.

Systems of Records Number(s) (SORN):

System Number:	System Name:
22.034	Individual Returns Files, Adjustments and Miscellaneous Documents
22.054	Subsidiary Accounting Files - Treasury/IRS
22.060	Automated Non-Master File (ANMF) - Treasury/IRS
22.062	Electronic Filing Records - Treasury/IRS
24.030	CADE Individual Master File (IMF) - Treasury/IRS
24.046	CADE Business Master File (BMF) - Treasury/IRS
26.020	Taxpayer Delinquency Investigation (TDI) Files Treasury/IRS
34.037	IRS Audit Trail and Security Records System – Treasury/IRS
42.008	Audit Information Management System (AIMS) – Treasury/IRS
42.021	Compliance Programs and Projects Files - Treasury/IRS

Data in the System

1. Describe the information (data elements and fields) available in the system in the following categories:

A. Taxpayer data elements and fields are available by **database type** in a number of files collectively described as a *data dictionary*. Each *data dictionary* is located on the CDW Website, under the **Database Library** webpage.

B. Employee: SEID, Name, [a]ol5081 password, and CDW account/logon.

C. Audit trail: Online 5081 registration validation as IRS employee/contractor to obtain password. Password delivered via secure email. Requestor must sign indicating receipt of password, and is secured by security officer. Requested applications are associated per user[b].

D. Other: Contractor(s) with root-level [c]access to build, develop, and maintain the CDW system, as well as its software and database models.

2. Describe/identify which data elements are obtained from files, databases, individuals, or any other sources.

A. IRS Files and Databases Used:
- ACS Action Codes and Narratives History
- AIMS – Audit Information Management System
- ARDI – Account Receivable Dollar Inventory System

- AUR – Automated Underreporter
- BMF – Business Master File
- BRTF – Business Returns Transaction File
- CRITS – Compliance Research Initiative Tracking System
- DI Data – Desktop Integration narratives from the field
- DM-1 File – Data Master-1 File
- EITC – Earned Income Tax Credit
- EOAD – Examination Operational Automation Database
- ERIS – Enforcement Revenue Information System
- IMF Bal Due Transactions History - Balance Due Modules
- IMF CCNIP/Return Delinquency – Taxpayer Nonfiler Categories
- IMF – Individual Master File
- IRTF – Individual Returns Transaction File
- IRMF – Information Returns Master File
- IRMF Treasury Department (TD) Form 90-22.1
- IRMF Form 1099-Original Issue Discount (OID)
- NRP – National Research Program

B. Taxpayer: TINs, Financial Institutions

C. Employee: TINs, SEID, Field-related ID's, Name, ol5081 password, data parameters, and CDW account/login

D. Federal Agency: Census Bureau Data

E. Third Party Source: US Postal Service Zip Codes (Files)

3. Is each data item required for the business purpose of the system? Explain.
Yes. Data in CDW are used to support a wide range of ad-hoc (that is, as the need arises; for example requests from the Commissioner) and unpredictable research questions, including those involving prediction, simulation, and optimization. As requests are on-going, some requests provide the opportunity to increase the amount/type of data elements/sets placed on CDW, thereby increasing aggregate data elements.

4. How will each data item be verified for accuracy, timeliness, and completeness?
The multiple databases on CDW contain one or more tables with data elements common across the databases. The tables are frequently matched across databases for inaccuracies that can occur during the extraction transformation or loading process. If inaccuracies are discovered to be a reflection of the source data, we do not attempt to correct these; but we do notify the system owner.

5. Is there another source for the data? Explain how that source is or is not used.
Sure. Individual datasets are available from the respective source owners as flat files, lacking the capability of analytical processing. CDW is the sole source storehouse of over 10 years worth of data that transforms data, derives new data elements, and validates and reorganizes the design and layout in a format conducive to analysis.

6. Generally, how will data be retrieved by the user?
Data is typically retrieved via third-party client tools, such as SQL, SAS, SPSS, and Hyperion.

7. Is the data retrievable by a personal identifier such as name, SSN, or other unique identifier?
Yes. Access to personal identifiers requires Executive-level approval, and is based on the business need established by a director level approved project plan; and specific instances, such as those stated in IRC 6103, Confidentiality and Disclosure of Tax Records.

Access to the Data

8. Who will have access to the data in the system (Users, Managers, System Administrators, Developers, Others)?
IRS employees, contractors, managers, and executives with a need to know, proper approval(s), evidence of security training completion, and who meet all applicable policies, standards, investigations, procedures, and safeguards in place.

9. How is access to the data by a user determined and by whom?
Requests for access to CDW are made through the Online Form 5081, and approved by the requestor's manager, security representative, and functional application manager. The approval process continues to the CDW Project Manager or designee, and then reviewed by the BSP/Security PMO. External requestors (such as contractors, Treasury, GAO et al) must use the paper Form 5081, and obtain approvals from the COTR, security representative, CDW Project Manager, Director, and the DAA. The BSP/Security PMO reviews the documentation, granting approvals and authorization based on signatory acknowledgement and agreement to practice the rules of system use, and to ensure security-related training has been completed and documented. The level of data access is limited to that which is specified in the approved proposal request and those established through assigned rights and privileges.

10. Do other IRS systems provide, receive, or share data in the system? If YES, list the system(s) and describe which data is shared.
CDW is a stand-alone system that does not interface with other systems. Data loads and queries are performed manually, and are primarily for IRS use only, except where exempted per IRC 6103, and related permissions and guidance appropriated by the Secretary of Treasury. See list in item # 2 above. IRS shares data that is necessary, on a need-to-know; available at the time of the request; and, when the requestor agrees to be accountable to policies, standards, investigations, procedures, and safeguards mentioned elsewhere in this document.

11. Have the IRS systems described in Item 10 received an approved Security Certification and Privacy Impact Assessment?
Yes, to the best of our knowledge.

12. Will other agencies provide, receive, or share data in any form with this system?
Yes. Congress, Treasury, GAO, TIGTA, and OMB are some of the agencies that may receive aggregate data, as they are entitled under IRC Section 6103, and all other applicable policies, standards, procedures, and safeguards of the Secretary of Treasury. (Such as, Computer Security Act of 1977, Treasury Regulation TDP 71-10, Federal Regulation Title 26 et al)

Administrative Controls of Data

13. What are the procedures for eliminating the data at the end of the retention period?
CDW records have yet to be scheduled for disposition as defined in IRM 1.15.27 for Compliance Research. Currently, the mixed data content of CDW has a range of dispositions up to and including PERMANENT records. In addition, these records contain information subject to the Disclosure

limitations of Section 6103 of the IRC. The IRS and NARA have yet to agree on specific procedures for purging the system of temporary records of the exact processes for secure transmission of the data between the two parties (The SF 115 is expected to be approved on or before January 1, 2010, and will reference the resolution of these issues).

14. Will this system use technology in a new way?
No

15. Will this system be used to identify or locate individuals or groups? If so, describe the business purpose for this capability.
Yes. CDW is used to perform research studies that may identify or predict taxpayer characteristics, for example, to measure noncompliance, evaluate the impact of program or policy changes on specific groups of taxpayers, and develop workload models to optimize the use of resources. These aggregate groups could be identified as a grouping. Recent causes for this capability include locating and identifying taxpayers - affected by Hurricane Katrina, Tsunami's, hurricanes, and tornado's - by geographical location to flag these accounts for deferment.

16. Will this system provide the capability to monitor individuals or groups? If yes, describe the business purpose for this capability and the controls established to prevent unauthorized monitoring.
Yes. CDW has tracking capability. By assigning a unique identifier to each TIN, we create masked TINs and in this way, we can track group-level data across multiple years. Other business purposes would include impact analyses on program changes, taxpayer behavior, predictive modeling, longitudinal surveys, and other research needs. Controls established to prevent unauthorized access and/or monitoring are based on permissions assigned and properly approved via the Form 5081 application, as well as annual recertification for security, privacy, disclosure, SBU, OUO, and other personally identifiable information.

17. Can use of the system allow IRS to treat taxpayers, employees, or others, differently?
Yes. CDW provides data to research analysts for purposes of estimation, prediction, simulation, optimization, and other statistical activities, some of which may result in new methods or approaches to classifying taxpayers on the basis of risk (for example, DIF scores), and allocating workload through compliance functions based on such methods or approaches.

18. Does the system ensure "due process" by allowing affected parties to respond to any negative determination, prior to final action?
Not applicable. CDW is read-only data. No capability exists to alter data. All data sets are derived from IMF or BMF Masterfile data. Should the taxpayer file subsequent corrected returns, or make any other adjustments, the resulting return data is eventually appended to the CDW system, but never replaced. CDW users do not make account decisions on individual taxpayers; instead, CDW provides the analytical processes used to make program/policy/treatment decisions, based on the population data.

19. If the system is web-based, does it use persistent cookies or other tracking devices to identify web visitors?
Not applicable.

View other PIAs on IRS.gov

[a] Solaris keeps accounts passwords in the "/etc/shadow" file, where the passwords are encrypted and stored as an encrypted string in this file. Each text password by default is encrypted by use of the "crypt (3c)" Solaris/Unix system call, which is generated by an algorithm using a "one-rotor machine" with "256-element rotor", which is technical jargon describing the encryption process used. Also, it is kept on the system, so that the server can be deployed and run independently. It also has the capability to interoperate via other protocols and authentication standards and technology.

[b] Information collected for audit trail includes, but is not limited to: all queries made by all users; all users identified by logon; no current process in place to use this information to deter any activities; however, users have read-only access

[c] Access levels are controlled through permissions established via password control; access rights and privileges are determined on an as-needed basis, as identified in the request for approval process.